DE GRUYTER
OPEN

Intracluster Homogeneity Selection Problem in a Business Survey

Berislav Žmuk
Faculty of Economics and Business, University of Zagreb, Croatia

Abstract

Background: In the cluster sampling approach many parameters have influence on lowering the survey costs and one of the most important is the intracluster homogeneity. **Objectives:** The goal of the paper is to find the most optimal value of intracluster homogeneity in case when two or more questions or variables have a key role in the research. **Methods/Approach:** Five key variables have been selected from a business survey conducted in Croatia and results for the two-stage cluster sampling design approach were simulated. The calculated intracluster homogeneity values were compared among all the five observed questions and survey costs and precision levels were inspected. **Results:** In the new cluster sampling design, for the fixed precision level, the lowest survey costs would be achieved by using the intracluster homogeneity value which is the closest to the average intracluster homogeneity value among all the key questions. Similar results were obtained when survey costs were held fixed. **Conclusions:** If there is more than one key question in the survey, then the best solution would be to use an average intracluster homogeneity value. However, one should notice that in that case minimum survey costs would not be reached, but the precision levels would increase at all key questions.

Keywords: business survey, cluster sampling, complex survey sampling design, design effect, key survey question, rate of homogeneity, survey costs

JEL classification: C83
Paper type: Research article

Received: Feb 02, 2016
Accepted: September 15, 2016

Acknowledgments: This work has been fully supported by the Croatian Science Foundation within the project STRENGTHS (project no. 9402).

Citation: Žmuk, B. (2016), "Intracluster Homogeneity Selection Problem in a Business Survey", Business Systems Research, Vol. 7, No. 2, pp. 91-103.
DOI: 10.1515/bsrj-2016-0015

Introduction

Nowadays, survey costs are becoming a more and more important parameter of a survey (Gonzalez, Eltinge, 2010, Krosnick et al., 2015). However, there is a sparse literature on survey costs (Karr, Last, 2006). In order to reduce survey other parameters, like precision and quality of the research, are often purposely disregarded and sacrificed (Groves, 1989, Schonlau, Fricker, Elliott, 2002). Different methods of data collection are developed to reduce survey costs (Groves et al., 2004). In order to reduce costs even more, in some cases it is justified to mix data collection methods (de Leeuw, 2005). Incentives initially do increase survey costs, but because they also increase response rates, at the end they could lead to deceased overall survey costs (Bricker, 2014).

In order to reduce survey costs, a researcher could choose a different data collection mode and/or different sampling design (Humphreys, 1979, Dillman, 1991, Groves, Heeringa, 2006). In this paper survey costs in cluster sampling design are investigated because this design often lead to the lowest survey costs under the same or similar parameters of a research (Daniel, 2012). Still, the lowest survey costs can be achieved only if an optimal balance of the number of clusters and their size is found (van Breukelen, Candel, 2012). Furthermore, the number of clusters and their size highly depends on the value of intracluster homogeneity. The intracluster homogeneity, which is estimated by rate of homogeneity (roh), measures the tendency of elements within a cluster to be correlated among themselves in comparison to the values of a variable for elements outside the cluster (Groves et al., 2004). Consequently, the intracluster homogeneity has an important role in survey costs.

The intracluster homogeneity is usually unknown and it is approximated by using rate of homogeneity from previous surveys which are very similar to the survey which is in plan to be conducted. The problem of finding similar surveys here is not going to be analysed, but the problem of finding right intracluster homogeneity value is going to be observed. Žmuk (2015b) has shown that lower survey costs are achieved when the intracluster homogeneity is lower. However, in the analysis he assumed that only one question in the survey was a target or key question. Consequently, only one intracluster homogeneity value is obtained and only this one value determines the number of clusters and their size. The problem arises when in the survey are more than just one key questions or key variables. Obviously, each variable has different intracluster homogeneity value. So, the main research question of this paper is: which intracluster homogeneity value in that case should be used to determine the number of clusters and their size? In order to give an answer to the research question survey costs and desired precision level of estimate are going to be taken into account.

The paper is outlined as followed. After the introduction part of the paper, cluster sampling characteristics are given in the second part of the paper. Methodology and survey data are which are going to be used in the analysis are presented in the third part. In the fourth part optimal intracluster homogeneity is calculated and analysed. The conclusions are provided in the final, fifth, part of the paper.

Cluster sampling methodology

There are two main reasons why a research would rather prefer cluster sampling than other methods of sampling (Levy, Lemeshow, 2008). The first reason why should cluster sampling be used is when a sampling frame for the whole population is not available. The costs of making complete sampling frame, which includes all the elements of the population that are under the study, could be very high. Also, sometimes there is needed a lot of time to complete the sampling frame. The consequence is that this sampling frame could not be more useful because of changes that happen in the population. The second reason for preferring cluster sampling is when the observed population is highly geographically dispersed. In that case cluster sampling lead to lower travelling costs.

In the cluster sampling it is assumed that the observed population can be divided into a number of certain nonoverlapping subpopulations or clusters (Bethlehem, 2009). If only a certain number of clusters are selected by using a sampling design then it is a case of one-stage sampling. On the other hand, if not all elements from sampled clusters are not selected, which means that there is another selection process within selected cluster, then it is a case of two-stage sampling.

Let it be assumed that there are overall *N* population elements in the sampling frame and that all *N* elements are eligible. The *N* population elements can be distributed among *A* clusters. So, the total number of clusters is equal to *A*. Each of formed clusters has *B* population elements. Obviously, every cluster is usually of different size or it has usually different number of population elements. In the one-stage cluster sampling design a certain number of clusters is selected and all elements within selected clusters are sampled. In the two-stage cluster sampling design, after selection of a certain number of clusters, population elements within selected clusters are sampled. In most cases different numbers of population elements within selected clusters are selected. Because of that the sample size in two-stage cluster sampling is given as:

$$n = a \cdot \bar{b}\,, \tag{1}$$

where n is the total sample size, a is the total number of selected clusters in the first stage and it is assumed that the total number of selected cluster is lower than the total number of cluster ($a < A$), $\bar{b}$ is the average number of selected elements in selected clusters calculated as overall number of selected elements in all selected clusters divided by the number of selected clusters.

In case of the two-stage cluster sampling design the overall mean statistics is equal to:

$$\bar{y} = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b} y_{\alpha\beta}}{a\bar{b}}\,, \tag{2}$$

whereas the sampling variance of the overall mean statistics is calculated as:

$$\mathrm{var}(\bar{y}) = \frac{1-f}{a\bar{b}^2(a-1)} \cdot \left[\sum_{\alpha=1}^{a} y_\alpha^2 - \frac{\left(\sum_{\alpha=1}^{a} y_\alpha\right)^2}{a} \right] = \frac{1-f}{a\bar{b}^2} \cdot s_a^2\,, \tag{3}$$

where $\bar{y}$ is the mean of the observed variable, $\alpha = 1,2,...,a$ are clusters in the sample, $\beta = 1,2,...,b$ are elements within cluster α, $y_{\alpha\beta}$ is the variable value of the element β in cluster α, a is the total number of selected clusters, $\bar{b}$ is the average number of selected elements in selected clusters, f is the sampling rate, s_a^2 is the between cluster variance.

The main disadvantage of cluster sampling is that the standard errors of estimates obtained from this design are usually higher than at other sampling designs (Levy, Lemeshow, 2008). The reason for that is that elements within a cluster are often homogeneous with respect to many characteristics but heterogeneous with elements in other clusters. Because of that Kish (1995) has introduced a measure which compares sampling variances of a complex sampling design and simple random sampling design. This measure is called "design effect" and in case of cluster sampling design it is calculated as follow:

$$deff = \frac{\mathrm{var}(\bar{y})_{CLUSTER}}{\mathrm{var}(\bar{y})_{SRS}}, \quad (4)$$

where $deff$ is the design effect, $\mathrm{var}(\bar{y})_{CLUSTER}$ is the sampling variance of mean in the cluster sampling design, $\mathrm{var}(\bar{y})_{SRS}$ is the sampling variance of mean in the simple random sampling design.

The intracluster homogeneity is defined as a measure of the homogeneity of the elements within clusters. Usually it is unknown and it must be estimated as the rate of homogeneity. In order to be able to calculate rate of homogeneity data from previous similar research are necessary. The rate of homogeneity is given as:

$$roh = \frac{deff - 1}{\bar{b} - 1}, \quad (5)$$

where roh is the rate of homogeneity, $deff$ is the design effect, $\bar{b}$ is the average number of selected elements in selected clusters. If the complete homogeneity within clusters is achieved, rate of homogeneity would be equal to 1. On the other hand, the maximum heterogeneity within clusters would result in rate of homogeneity of $-1/(\bar{b} - 1)$.

Costs in cluster sampling design include fixed or administrative costs and field costs. The value of field costs depends on the number of selected clusters and the number of selected elements within the clusters. Therefore, the function of costs in cluster sampling design is:

$$C = C_0 + a \cdot c_a + a \cdot \bar{b} \cdot c_b, \quad (6)$$

where C are total survey costs, C_0 are fixed costs, a is the total number of selected clusters, c_a is the cost per cluster, $\bar{b}$ is the average number of selected elements in selected clusters, c_b is the cost per element within a cluster. If the survey budget is limited and in forward known, the optimal number of clusters and their size can be obtained by using the Lagrange multiplier or the Cauchy-Schwartz inequality (Cochran, 1977, Varberg, Purcell, 1997).

Methodology and survey data

In order to inspect which intracluster homogeneity value should be used to determine the number of clusters and their size in case of more than one key variable, variables and data from a business survey in Croatia are used. In the business survey provided their attitudes towards statistical methods and answered how often they use certain statistical methods in their businesses (Žmuk, 2015a). Simple random sampling design was used as a sampling design in the survey. Still, after conducted survey enterprises were stratified according their size, main activity and legal form (Žmuk, 2013).

The survey population consisted of 58,954 Croatian enterprises which have been doing business at least since 2011. Due to sampling frame limitations, the sampling population was consisted of 26,186 enterprises. The enterprises got invitation for participation in the survey by e-mail in October 2012. In the e-mail a unique hyperlink to the web questionnaire was provided also. Overall 667 enterprises successfully participated and fulfilled the questionnaire by the middle of February 2013. On that way the Response rate 1 of 1.13% was achieved (American Association for Public Opinion Research, 2015).

For the purpose of the analysis in this paper, five questions from the survey are sorted out and declared to be the key questions. According to characteristics of a key question needed sample size for achieving certain precision level can be determined. In the paper this five questions are going to be first analysed separately and then altogether. At all five key questions parameter of interest is proportion. In accordance with that some adjustments must have been done to get only two possible answers, positive “Yes” or negative “No”, to each key question. So, answers “I don’t know” are removed from the analysis. Onwards, depending on whether an enterprise uses statistical methods in their business or not it has got an option to answer different set of questions. All this adjustments and a filter question lead to different number of answers at each key variable. The key questions and their basic survey results are shown in Table 1.

Table 1
Survey key questions and their basic survey results

Key question	Positive answers ("Yes")	Negative answers ("No")	Total answers	Proportion of "Yes" answers	Simple random sampling variance
Q1. Do you use statistical methods in your business?	237	430	667	0.3553	0.000344
Q2. Are you using statistical methods as a support in decision making?	213	11	224	0.9509	0.000209
Q3. Are you investing in statistical software use?	102	100	202	0.5050	0.001244
Q4. Has statistical methods use improved your business results?	186	16	202	0.9208	0.000363
Q5. Statistical methods are not used in your enterprise because employees are not well known with statistical methods in general?	210	153	363	0.5785	0.000674

Note: In order to calculate simple random sampling variances the sampling rate lower than 0.05 was assumed.
Source: author’s calculation.

According to the results provided in Table 1, the most answers enterprises provided on the first key question Q1 (667 answers) whereas the least answers were provided on the key questions Q3 and Q4 (202 answers). The proportions of positive answers differ between the key questions from 0.3553 at Q1 to 0.9509 at Q2. Consequently, there is also difference in simple random sampling variances at the key questions. All these differences in the further analysis should result in different needed sample sizes and in different survey costs.

In order to inspect problem of selecting the most appropriate or optimal intracluster homogeneity value when there is more than one key question or variable, intracluster homogeneity values for each of the five key variables are going to be calculated. In order to estimate intracluster homogeneity by rate of homogeneity, the rates of homogeneity are calculated by assuming that previously described survey was conducted by using two-stage cluster sampling design with probabilities proportionate to the size.

First, there are going to be calculated cluster sampling variances for each of key variables. The roles of clusters are going to have counties of the Republic of Croatia. There are 20 counties plus the City of Zagreb and so 21 clusters of enterprises are formed. Enterprises are associated with clusters according to place of their headquarters. In order to obey two-stage cluster sampling design characteristics, it is assumed that there are more than 21 clusters.

In the next step cluster sampling variances and simple random sampling variances are compared and design effects are calculated.

After that rates of homogeneity are calculated for each key question separately. In the further analysis the values of survey costs, sample sizes and precision levels for the calculated rates of homogeneity are observed and compared.

Selection of optimal intracluster homogeneity value

Instead of simple random sampling design in the observed survey about statistical methods use in Croatian enterprises, it is assumed that two-stage cluster sampling design was applied. There are 20 counties plus the City of Zagreb in Croatia. Consequently it is assumed that there are selected 21 clusters. Because the number of answers is different across the five key questions, and because enterprises are classified into the clusters according place of theirs headquarters, the number of elements or enterprises per cluster is very different. So, according to Tables 2-6, the minimal cluster size was one, and the maximum size was 249. In Tables 2-6 are separately given basic cluster sampling results for the five key questions according to the clusters.

Table 2

Basic cluster sampling results for the 1st key question, a=21 clusters, n=667 enterprises

Counties (clusters)	Q1. Do you use statistical methods in your business?				
	"Yes" answers	"No" answers	Total answers	Proportion of "Yes" answers	Cluster variance
Bjelovar-Bilogora	4	8	12	0.3333	0.0202
City of Zagreb	75	174	249	0.3012	0.0008
Dubrovnik-Neretva	3	10	13	0.2308	0.0148
Istria	17	27	44	0.3864	0.0055
Karlovac	2	4	6	0.3333	0.0444
Koprivnica-Križevci	4	11	15	0.2667	0.0140
Krapina-Zagorje	6	12	18	0.3333	0.0131
Lika-Senj	2	1	3	0.6667	0.1111
Međimurje	7	10	17	0.4118	0.0151
Osijek-Baranja	12	14	26	0.4615	0.0099
Požega-Slavonia	1	4	5	0.2000	0.0400
Primorje-Gorski kotar	25	39	64	0.3906	0.0038
Sisak-Moslavina	6	7	13	0.4615	0.0207
Slavonski Brod-Posavina	4	5	9	0.4444	0.0309
Split-Dalmatia	16	33	49	0.3265	0.0046
Šibenik-Knin	6	11	17	0.3529	0.0143
Varaždin	13	19	32	0.4063	0.0078
Virovitica-Podravina	2	7	9	0.2222	0.0216
Vukovar-Sirmium	3	10	13	0.2308	0.0148
Zadar	3	4	7	0.4286	0.0408
Zagreb	26	20	46	0.5652	0.0055
Total	237	430	667	-----	-----

Source: author's calculation.

Table 3
Basic cluster sampling results for the 2nd key question, a=21 clusters, n=224 enterprises

Counties (clusters)	Q2. Are you using statistical methods as a support in decision making?				
	"Yes" answers	"No" answers	Total answers	Proportion of "Yes" answers	Cluster variance
Bjelovar-Bilogora	3	0	3	1.0000	0.0000
City of Zagreb	66	5	71	0.9296	0.0009
Dubrovnik-Neretva	3	0	3	1.0000	0.0000
Istria	17	0	17	1.0000	0.0000
Karlovac	2	0	2	1.0000	0.0000
Koprivnica-Križevci	3	0	3	1.0000	0.0000
Krapina-Zagorje	6	0	6	1.0000	0.0000
Lika-Senj	2	0	2	1.0000	0.0000
Međimurje	7	0	7	1.0000	0.0000
Osijek-Baranja	12	0	12	1.0000	0.0000
Požega-Slavonia	1	0	1	1.0000	-----
Primorje-Gorski kotar	21	2	23	0.9130	0.0036
Sisak-Moslavina	5	1	6	0.8333	0.0278
Slavonski Brod-Posavina	4	0	4	1.0000	0.0000
Split-Dalmatia	15	0	15	1.0000	0.0000
Šibenik-Knin	6	0	6	1.0000	0.0000
Varaždin	9	2	11	0.8182	0.0149
Virovitica-Podravina	1	0	1	1.0000	-----
Vukovar-Sirmium	3	0	3	1.0000	0.0000
Zadar	3	0	3	1.0000	0.0000
Zagreb	24	1	25	0.9600	0.0016
Total	213	11	224	-----	-----

Source: author's calculation.

Table 4
Basic cluster sampling results for the 3rd key question, a=21 clusters, n=202 enterprises

Counties (clusters)	Q3. Are you investing in statistical software use?				
	"Yes" answers	"No" answers	Total answers	Proportion of "Yes" answers	Cluster variance
Bjelovar-Bilogora	2	1	3	0.6667	0.1111
City of Zagreb	29	31	60	0.4833	0.0042
Dubrovnik-Neretva	1	1	2	0.5000	0.2500
Istria	12	5	17	0.7059	0.0130
Karlovac	1	0	1	1.0000	-----
Koprivnica-Križevci	2	0	2	1.0000	0.0000
Krapina-Zagorje	2	4	6	0.3333	0.0444
Lika-Senj	1	0	1	1.0000	-----
Međimurje	2	5	7	0.2857	0.0340
Osijek-Baranja	6	5	11	0.5455	0.0248
Požega-Slavonia	0	1	1	0.0000	-----
Primorje-Gorski kotar	12	11	23	0.5217	0.0113
Sisak-Moslavina	3	3	6	0.5000	0.0500
Slavonski Brod-Posavina	1	3	4	0.2500	0.0625
Split-Dalmatia	7	9	16	0.4375	0.0164
Šibenik-Knin	1	5	6	0.1667	0.0278
Varaždin	5	6	11	0.4545	0.0248
Virovitica-Podravina	1	0	1	1.0000	-----
Vukovar-Sirmium	2	1	3	0.6667	0.1111
Zadar	0	1	1	0.0000	-----
Zagreb	12	8	20	0.6000	0.0126
Total	102	100	202	-----	-----

Source: author's calculation.

Table 5
Basic cluster sampling results for the 4th key question, a=21 clusters, n=202 enterprises

Counties (clusters)	Q4. Has statistical methods use improved your business results?				
	"Yes" answers	"No" answers	Total answers	Proportion of "Yes" answers	Cluster variance
Bjelovar-Bilogora	3	0	3	1.0000	0.0000
City of Zagreb	55	9	64	0.8594	0.0019
Dubrovnik-Neretva	3	0	3	1.0000	0.0000
Istria	13	2	15	0.8667	0.0083
Karlovac	2	0	2	1.0000	0.0000
Koprivnica-Križevci	4	0	4	1.0000	0.0000
Krapina-Zagorje	5	1	6	0.8333	0.0278
Lika-Senj	2	0	2	1.0000	0.0000
Međimurje	7	0	7	1.0000	0.0000
Osijek-Baranja	10	1	11	0.9091	0.0083
Požega-Slavonia	1	0	1	1.0000	-----
Primorje-Gorski kotar	19	1	20	0.9500	0.0025
Sisak-Moslavina	6	0	6	1.0000	0.0000
Slavonski Brod-Posavina	4	0	4	1.0000	0.0000
Split-Dalmatia	14	0	14	1.0000	0.0000
Šibenik-Knin	6	0	6	1.0000	0.0000
Varaždin	7	1	8	0.8750	0.0156
Virovitica-Podravina	1	0	1	1.0000	-----
Vukovar-Sirmium	2	0	2	1.0000	0.0000
Zadar	1	1	2	0.5000	0.2500
Zagreb	21	0	21	1.0000	0.0000
Total	186	16	202	-----	-----

Source: author's calculation.

Table 6
Basic cluster sampling results for the 5th key question, a=21 clusters, n=363 enterprises

Counties (clusters)	Q5. Statistical methods are not used in your enterprise because employees are not well known with statistical methods in general?				
	"Yes" answers	"No" answers	Total answers	Proportion of "Yes" answers	Cluster variance
Bjelovar-Bilogora	2	4	6	0.3333	0.0444
City of Zagreb	82	63	145	0.5655	0.0017
Dubrovnik-Neretva	6	1	7	0.8571	0.0204
Istria	18	6	24	0.7500	0.0082
Karlovac	3	1	4	0.7500	0.0625
Koprivnica-Križevci	4	6	10	0.4000	0.0267
Krapina-Zagorje	4	6	10	0.4000	0.0267
Lika-Senj	0	1	1	0.0000	----
Međimurje	5	3	8	0.6250	0.0335
Osijek-Baranja	9	4	13	0.6923	0.0178
Požega-Slavonia	1	2	3	0.3333	0.1111
Primorje-Gorski kotar	19	14	33	0.5758	0.0076
Sisak-Moslavina	3	3	6	0.5000	0.0500
Slavonski Brod-Posavina	2	2	4	0.5000	0.0833
Split-Dalmatia	15	13	28	0.5357	0.0092
Šibenik-Knin	7	4	11	0.6364	0.0231
Varaždin	10	7	17	0.5882	0.0151
Virovitica-Podravina	4	2	6	0.6667	0.0444
Vukovar-Sirmium	3	4	7	0.4286	0.0408
Zadar	2	1	3	0.6667	0.1111
Zagreb	11	6	17	0.6471	0.0143
Total	210	153	363	-----	-----

Source: author's calculation.

In order to calculate cluster variances in Tables 2-6 the sampling rates lower than 0.05 were assumed. If a cluster was consisted of only one element or just one enterprise, the cluster variance could not be calculated. In Table 2-6 cluster variances for each cluster are provided but the overall cluster sampling variance must be calculated in the next step. Because the clusters are of unequal sizes, the ratio approach to calculation of cluster sampling variance must be used (Kish, 1995). Consequently, the cluster sampling variances for each of the five key questions were calculated by using following equation:

$$\operatorname{var}(r) = \frac{1-f}{\left(\sum_{\alpha=1}^{a} b_\alpha\right)^2} \cdot \frac{a}{a-1} \cdot \left[\sum_{\alpha=1}^{a} y_\alpha^2 + r^2 \cdot \sum_{\alpha=1}^{a} b_\alpha^2 - 2 \cdot r \cdot \sum_{\alpha=1}^{a} y_\alpha b_\alpha\right], \tag{7}$$

where r is the ratio (proportion), f is the sampling rate, b_α is the number of selected elements in the selected cluster α, $\alpha = 1,2,...,a$ are clusters in the sample, a is the total number of selected clusters, y_α is the number of elements with the chosen characteristic in cluster α. Again, it is assumed that the sampling rates at the five key questions are negligible. The cluster sampling variances are given in Table 7.

Table 7
Rates of homogeneity for the five key questions

Counties (clusters)	Key questions				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Number of clusters	21	21	21	21	21
Average number of elements in clusters	31.76	10.67	9.62	9.62	17.29
Total sample size	667	224	202	202	363
Simple random sampling variance	0.00034	0.00021	0.00124	0.00036	0.00067
Cluster sampling variance	0.00073	0.00016	0.00077	0.00059	0.00033
Design effect	2.1131	0.7773	0.6219	1.6128	0.4911
Rate of homogeneity	0.0362	-0.0230	-0.0439	0.0711	-0.0313

Source: author's calculation.

After cluster sampling variances, design effects for all five key questions are calculated and are shown in Table 7. At the first, Q1, and the fourth, Q4, key questions cluster sampling variance is higher than simple random sampling variance. Consequently, the design effects at these two questions are higher than one. At the other three key questions cluster sampling variance is lower than simple random sampling variance. This situation is not usual when cluster sampling as a complex design is used but it can happen.

When the design effect is known, then the calculation of rate of homogeneity is straightforward. Rates of homogeneity values for the five key questions are shown in the last row in Table 7. The maximum rate of homogeneity was achieved at the fourth key question (roh=0.0711) whereas the lowest rate of homogeneity is at the third key question (roh=-0.0439). Obviously all calculated rates of homogeneity are different. So, the question is which rate of homogeneity, which is an estimation of intracluster homogeneity, should be used in the new survey to determine number of clusters, cluster sizes and sample size? If there was just one key variable the answer is very easy but here it is unclear which the best or optimal solution is.

In order to examine which intracluster homogeneity value from the five provided should be used as optimal one, two different approaches are going to be used. In the first approach the lowest survey costs criteria for selection of optimal intracluster homogeneity are going to be used. On the other hand, the required precision of an estimate is going to be used as a criterion for intracluster homogeneity selection. In both approaches it is estimated that costs per cluster are €500 and the costs per element within a cluster are €25. Furthermore, the confidence level of 95% is used. In the first approach, where survey costs are calculated, precision as confidence interval or margin of error of 5% or 0.05 is defined. On the other hand, in the second approach, where precision is calculated, survey costs of €30,000 are given. In the first approach number of cluster is estimated by using following equation:

$$a = \frac{p \cdot (1-p) \cdot z^2}{\bar{b} \cdot e^2} \cdot \left[roh \cdot (\bar{b} - 1) + 1\right], \tag{8}$$

where a is the total number of selected clusters, p is the expected proportion used from the previous research, z is the value from the normal distribution, based on the desired level of confidence, $\bar{b}$ is the average number of selected elements in selected clusters, e is the absolute value of the tolerated sampling variance which is based on the required precision, roh is the rate of homogeneity (Leon et al., 2014). In the both approaches it is assumed that the average number of selected elements in selected clusters is equal to 20. In Table 8 results for the first approach and in Table 9 results for the second approach are provided.

Table 8
Survey costs for the five key questions, results of the first approach

Statistics	Key questions				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Expected proportion	0.3553	0.9509	0.5050	0.9208	0.5785
Normal distribution value	1.96	1.96	1.96	1.96	1.96
Rate of homogeneity	0.0362	-0.0230	-0.0439	0.0711	-0.0313
Average number of elements per cluster	20	20	20	20	20
Tolerated sampling variance	0.05	0.05	0.05	0.05	0.05
Number of clusters	29.70	2.02	3.20	13.17	7.61
Final number of clusters	30	3	4	14	8
Sample size	600	60	80	280	160
Cost per cluster	500	500	500	500	500
Cost per element within a cluster	25	25	25	25	25
Total survey costs	30,000	3,000	4,000	14,000	8,000

Source: author's calculation.

According to the results from Table 8, the first key question requires the highest amount of survey costs (€30.000) for obtaining the same level of precision like other key questions. On the other hand, the second key question requires the lowest amount of survey costs (€3.000). If the situation from the survey costs is observed than the best solution would be to use parameters from the second key question in the new cluster sampling design. However, the sample size at the second key question is the lowest which would lead to lower precision level at other key questions. It has to be emphasized that rate of homogeneity at the second key question is neither the highest nor the lowest among the observed key questions. So, the optimal solution would not be to take either the highest or the lowest intracluster homogeneity value. The average rate of homogeneity for the five key questions is equal to 0.0018 and the rate of homogeneity of the second key question is the nearest to this value. This conclusion speaks in favour of using average rate of homogeneity of all key variables in the new cluster sampling design. By using of average rate of homogeneity the survey costs would rise because of increased sample size, but in the same time precision level at all key variables would rise also.

Table 9
Survey precision for the five key questions, results of the second approach

Statistics	Key questions				
	Q1	Q2	Q3	Q4	Q5
Total survey costs	30,000	30,000	30,000	30,000	30,000
Average number of elements per cluster	20	20	20	20	20
Cost per cluster	500	500	500	500	500
Cost per element within a cluster	25	25	25	25	25
Number of clusters	30	30	30	30	30
Sample size	600	600	600	600	600
Expected proportion	0.3553	0.9509	0.5050	0.9208	0.5785
Normal distribution value	1.96	1.96	1.96	1.96	1.96
Rate of homogeneity	0.0362	-0.0230	-0.0439	0.0711	-0.0313
Tolerated sampling variance	0.0497	0.0130	0.0163	0.0331	0.0252

Source: author's calculation.

According to the results in Table 9, the highest precision level for given survey costs is achieved at the second key question. On the other hand the lowest precision level seems to be at the first key question. These results are analogous to the results from Table 8 and confirm the connection between survey costs and precision level. Consequently, the same conclusion about the optimal intracluster homogeneity value as an average of intracluster homogeneity values at all key variables can be made as before.

Conclusion

Cluster sampling design is very popular among researchers because by its use considerable savings on survey costs can be made. Also, it is recommended sampling design when sampling frame is not perfect or of high quality. However, it has to be kept on mind that cluster sampling design usually has lower precision level in compare to the simple random sampling for the same sample size.

The very important parameter of cluster sampling design is intracluster homogeneity which measures the correlation of elements within a cluster in comparison to the elements in other clusters. Because intracluster homogeneity is not known, as its approximation rate of homogeneity is used. The rate of homogeneity is estimated based on previous similar surveys which had very similar key questions. Based on the rate of homogeneity the number of clusters, cluster size and sample size are determined in the new cluster sampling design. Because these parameters have significant role on the survey costs and precision level, the rate of homogeneity has to be carefully chosen.

If a research can declare only one question as very important or the key one from the questionnaire, rate of homogeneity is rather easy to calculate. However, the problem appears when there are more key questions. In the paper it was investigated which rate of homogeneity should be used if there are five key variables. In the research data from previous survey about statistical methods use in Croatian enterprises was used. Based on this data cluster sampling design was simulated and rates of homogeneity for each of five key variables were calculated. The results have shown that neither the lowest nor the highest rates of homogeneity can be observed as optimal ones. In fact, the lowest quality costs and the highest precision level was achieved when was used rate of homogeneity which was nearest to the average of all five observed rates of homogeneity. So, the optimal intracluster homogeneity value, which minimizes survey costs and maximizes precision level, could be the value which represents the average intracluster homogeneity value of all key questions.

The main limitation of the paper is that in the analysis are not used data from survey which was based on the cluster sampling design but this design was simulated. Furthermore, the calculated rates of homogeneity were quite similar and, what could be a bigger problem, all these values were very close to zero value. It would be of interest to investigate and to find optimal intracluster homogeneity value if these values at the key variables were more different. In addition, in further research more cases and different parameters of cluster sampling should be used to check theirs impact on the optimal intracluster homogeneity value.

References

1. American Association for Public Opinion Research (2015), "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys", available at http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf (27 January 2016)
2. Bethlehem, J. (2009). Applied Survey Methods: A Statistical Perspective, Hoboken, John Wiley & Sons.
3. Bricker, J. (2014). Survey Incentives, Survey Effort, and Survey Costs. FEDS Working Paper No. 2014-74, Washington, pp. 1-36.
4. Cochran, W. G. (1977). Sampling Techniques, New York, John Wiley and Sons.
5. Daniel, J. (2012). Sampling Essentials: Practical Guidelines for Making Sampling Choices, Thousand Oaks, California, SAGE Publications.
6. de Leeuw, E. D. (2005), "To Mix or Not to Mix Data Collection Modes in Surveys", The Journal of Official Statistics, Vol. 21, No. 2, pp. 233-255.
7. Dillman, D. A. (1991), "The Design and Administration of Mail Surveys", Annual Review of Sociology, Vol. 17, pp. 225-249.
8. Gonzalez, J. M., Eltinge, J. L. (2010), "Optimal Survey Design: A Review", available at http://www.bls.gov/osmr/pdf/st100270.pdf (27 January 2016)
9. Groves, R. M. (1989). Survey Errors and Survey Costs, New Jersey, John Wiley & Sons.
10. Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2004). Survey Methodology. Hoboken, New Jersey, John Wiley & Sons.
11. Groves, R. M., Heeringa, S. G. (2006), "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs", Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 169, No. 3, pp. 439-457.
12. Humphreys, C. P. (1979), "The cost of sample survey designs", Proceedings Section on Survey Research Methods, American Statistical Association, pp. 395-400.
13. Karr, A. F., Last, M. (2006), "Survey Costs: Workshop Report and White Paper", Technical Report Number 161, September 2006, National Institute of Statistical Sciences, pp. 1-22.
14. Kish, L. (1995). Survey Sampling, New York, John Wiley & Sons.
15. Krosnick, J. A., Presser, S., Fealing, K. H., Ruggles, S. (2015), "The Future of Survey Research: Challenges and Opportunities", available at http://www.nsf.gov/sbe/AC_Materials/The_Future_of_Survey_Research.pdf (27 January 2016)
16. Leon, E. A., Perez, A.M., Stevenson, M.A., Robiolo, B., Mattion, N., Seki, C., La Torre, J., Torres, A., Cosentino, B., Duffy, S. J. (2014), "Effectiveness of systematic

foot and mouth disease mass vaccination campaigns in Argentina", Scientific and Technical Review, Vol. 33, No. 3, pp. 917-926.
17. Levy, P. S., Lemeshow, S. (2008). Sampling of Populations: Methods and Applications, Hoboken, John Wiley & Sons.
18. Schonlau, M., Fricker, R. D., Elliott, M. N. (2002). Conducting Research Surveys via E-mail and the Web, Santa Monica, RAND Corporation.
19. van Breukelen, G. J., Candel, M. J. (2012), "Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient!", Journal of clinical epidomology, Vol. 65, No. 11, pp. 1212-1218.
20. Varberg, D., Purcell, E. J. (1997). Calculus, New Jersey, Prentice Hall.
21. Žmuk, B. (2013), "The Relevance of Statistical Methods Application to Business Performance of Enterprises", doctoral thesis, Zagreb, Faculty of Economics and Business.
22. Žmuk, B. (2015a), "Business Sample Survey Measurement on Statistical Thinking and Methods Adoption: the Case of Croatian Small Enterprises", Interdisciplinary description of complex systems, Vol. 13, No. 1, pp. 154-166.
23. Žmuk, B., (2015b), "The impact of intracluster homogeneity on the survey costs: A Croatian business survey", Proceedings of the 13th International Symposium on Operational Research SOR'15 (Editors: Zadnik Stirn, L., Žerovnik, J., Kljajić Borštnar, M., Drobne, S.), Bled, Slovenia, September 23-25, 2015, Ljubljana: Slovenian Society Informatika, Section for Operational Research, pp. 329-334.

About the author

Berislav Žmuk, PhD, graduated at the major Accounting, post-graduated Statistical Methods for Economic Analysis and Forecasting, and gained his PhD degree in Business Economics at Faculty of Economics and Business, University of Zagreb. Currently he is a Senior Assistant at the Department of Statistics, Faculty of Economics and Business, University of Zagreb where he teaches following subjects: Statistics, Business Statistics and Business Forecasting. In 2013, he successfully completed Sampling Program for Survey Statisticians (SPSS) at Survey Research Center (SRC), Institute for Social Research (ISR), and University of Michigan in Ann Arbor, Michigan, USA. In 2015, he completed several survey methodology courses (Introduction to Web Surveys, Introduction to Questionnaire Design, Mixed-Mode and Mixed-Device Surveys) at Gesis, Leibniz Institute for Social Research in Cologne, Germany. His main research fields include applications of statistics in business and economy, survey methodology and statistical quality control. The author can be contacted at **bzmuk@efzg.hr**